Semi-supervised feature extraction using independent factor analysis

Latifa Oukhellou, Etienne Côme, Patrice Aknin, Thierry Denoeux

Semi-supervised feature extraction using independent factor analysis

L. Oukhellou[b,c], E. Côme[a], P. Aknin[b], Th. Denœux[d]

[a] *SAMOS-MATISSE, Université Paris 1, 90, rue de Tolbiac, 75634 paris cedex 13, France*
[b] *French National Institute for Transport and Safety Research, 2 av. Malleret-Joinville, 94114 Arcueil Cedex, France*
[c] *Université Paris 12 - CERTES, 61 av. du Général de Gaulle, 94100 Créteil, France*
[d] *Heudiasyc, UTC - UMR CNRS 6599, B.P 20529, 60205 Compiègne - France*

Abstract

Dimensionality reduction can be efficiently achieved by generative latent variable models such as probabilistic principal component analysis (PPCA) or independent component analysis (ICA), aiming to extract a reduced set of variables (latent variables) from the original ones. In most cases, the learning of these methods is achieved within the unsupervised framework where only unlabeled samples are used. In this paper we investigate the possibility of estimating independent factor analysis model (IFA) and thus projecting original data onto a lower dimensional space, when prior knowledge on the cluster membership of some training samples is incorporated. In the basic IFA model, latent variables are only recovered from their linear observed mixtures (original features). Both the mapping matrix (assumed to be linear) and the latent variable densities (that are assumed to be mutually independent and generated according to mixtures of Gaussians) are learned from observed data. We propose to learn this model within semi-supervised framework where the likelihood of both labeled and unlabeled samples is maximized by a generalized expectation-maximization (GEM) algorithm. Experimental results on real data sets are provided to demonstrate the ability of our approach to find law dimensional manifold with good explanatory power.

Key words: Independent factor analysis, semi-supervised learning, mixture models, maximum likelihood, dimensionality reduction

1. Introduction

A considerable amount of research has been devoted to dimensionality reduction of multivariate data sets. The underlying motivation is that many data sets live on a subspace whose intrinsic dimensionality is lower than that of the original data space. Research in this area has employed either feature selection methods which directly select a subset of meaningful variables from the original ones, or feature extraction methods that aim to generate new features from the first input data. Whatever the chosen approach, the goal is to describe a large data set by a reduced number of variables that better capture the essential structure of the problem. These methods include linear approaches such as probabilistic principal component analysis (PPCA) [1, 2] Projection Pursuit (PP) [12, 13], Metric Muldimensional Scaling (MDS) [8], linear discriminant analysis (LDA) [11] on the one hand, and non linear methods such as random projection (RP) [14, 15], Kohonen's self-organizing maps [35, 32],... on the other hand. A detailed survey of many of these methods can be found in [9, 10].

The dimensionality reduction problem can also be formulated using a generative latent variable model which aims to describe observed variables (original features), in terms of smaller set of unobservable (or latent) variables. Depending on the assumption made on the latent and observed variable distributions, different kind of models can be distinguished such as Principal Component Analysis (PCA), Factor analysis (FA) [30, 31], and Independent Component Analysis (ICA) [16, 17, 29]. ICA has been applied to many different problems, including blind source separation, exploratory data analysis and feature extraction [34, 21]. In the feature extraction context, several authors used ICA to extract meaningful features for both regression and classification problems [37, 38]. This paper deals with a particular model of this family, recently proposed by [18, 19], and known as Independent Factor Analysis (IFA).

The generative model involved in IFA assumes that observed variables are generated by a linear mixture of independent and non Gaussian latent variables as in the ICA model. Furthermore, it considers that each individual latent variable has its own distribution, modeled by a mixture of Gaussians (MOG). The IFA model is often considered within an unsupervised learning framework. The model parameters and thus the latent variables are learned from the observed data only. Recent works have derived an approach for modeling class conditional densities based on IFA model [20]. In this paper, we propose an extension of the basic IFA model makes it possible to incorporate additional information on cluster membership of some training samples to estimate the IFA model. In this way, the learning of this model, and thus the dimensionality reduction can be handled in a semi-supervised learning framework.

The article is organized as follows. In Section 2, we review the independent factor analysis model and present how it can be

Email address: oukhellou@inrets.fr Tel: +33 1 45 92 56 58 (L. Oukhellou)

estimated by maximum likelihood in a noiseless setting. Section 3 focuses on the problem of semi-supervised learning of the IFA model where additional information on cluster membership of some samples will be incorporated. A generalized maximum likelihood criterion will be defined and the algorithm for its optimization also detailed. Experimental results showing the benefits of the proposed approach to achieve dimensionality reduction will then be given for real data sets. Conclusions are presented in Section 5.

2. Noiseless Independent Factor Analysis

2.1. Background on Independent Factor Analysis

IFA was introduced in [19, 18]. It originates from both ordinary factor analysis (FA) in applied statistics [27, 28] and Independent Component Analysis (ICA) in signal processing [16, 17]. IFA aims to recover independent latent variables from their observed linear mixtures. The latent variables are assumed to be mutually independent and non Gaussians. In the noiseless form that is used throughout this paper, the IFA model can be expressed as:

$$\mathbf{x} = A\,\mathbf{z}, \tag{1}$$

where A is a square matrix of size $S \times S$, $\mathbf{x}$ the random vector whose elements $(\mathbf{x}_1, \ldots, \mathbf{x}_S)$ are the mixtures and $\mathbf{z}$ the random vector whose elements $(\mathbf{z}_1, \ldots, \mathbf{z}_S)$ are the latent variables. Thanks to the noiseless setting, a deterministic relationship between the distributions of observed and latent variables can be expressed as:

$$f^{\mathcal{X}}(\mathbf{x}) = \frac{1}{|\det(A)|} f^{\mathcal{Z}}(A^{-1}\,\mathbf{x}), \tag{2}$$

Unlike the ICA model in which the probability density functions of the latent variables are fixed using prior knowledge or according to some indicator that allows switching between sub and super Gaussian densities [16], each latent variable density in the IFA is modeled as a mixture of normally distributed components (Mixture of Gaussians MOG) so that a wide class of densities can be approximated [19, 18, 7]:

$$f^{\mathcal{Z}_s}(z_s) = \sum_{k=1}^{K_s} \pi_k^s \varphi(z_s; \mu_k^s, \nu_k^s), \tag{3}$$

where $\varphi(.; \mu, \nu)$ denotes a univariate normal density function with mean μ and variance ν. Equation (3) means that each latent variable is described as mixture of K_s Gaussians with mean μ_k^s, variance ν_k^s and mixing proportions π_k^s.

Considering the graphical model of IFA shown in Figure 1, it can be seen that the IFA model provides two levels of interpretation corresponding to discrete and continuous latent variables. For each one of the S latent variables, the discrete latent variable encodes the cluster from which each sample is drawn.

The whole IFA model parameters can thus be summarized in a vector $\boldsymbol{\psi} = (W, \boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^S, \boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^S, \boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^S)$, with W the unmixing matrix ($W = A^{-1}$), $\boldsymbol{\pi}^j$ the vector of cluster proportions of source j which sum to 1, $\boldsymbol{\mu}^j$ and $\boldsymbol{\nu}^j$ the vectors of size K_j containing the means and the variances of each cluster.

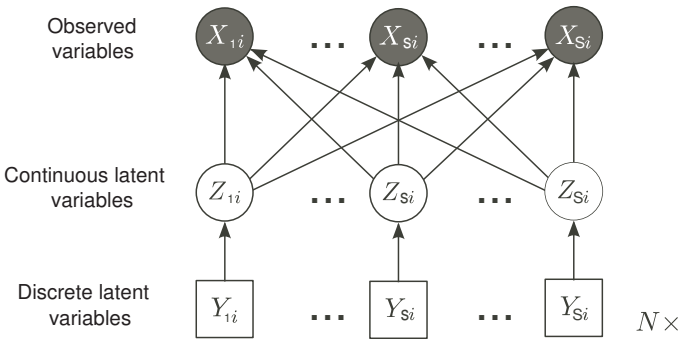

Figure 1: Graphical model for Independent Factor Analysis.

2.2. Parameter estimation in IFA

2.2.1. IFA and Maximum Likelihood

The learning problem associated with the IFA model consists in estimating both the unmixing matrix W and the MOG parameters from the observed variables alone. Considering an i.i.d random sample $\mathbf{X} = (x_1, ..., x_N)$ of size N and using Equation (2) under the latent variable independence hypothesis, the log-likelihood has the form:

$$\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}) = \sum_{i=1}^{N} \sum_{s=1}^{S} \log\left(f^{\mathcal{Z}_s}((W\mathbf{x}_i)_s)\right) + N \log(|\det(W)|), \tag{4}$$

By substituting the density distribution by its expression given in (3), the log-likelihood can be rewritten as:

$$\mathcal{L}(\boldsymbol{\psi}; \mathbf{X}) = N \log(|\det(W)|) + \sum_{i=1}^{N} \sum_{s=1}^{S} \log\left(\sum_{k=1}^{K_s} \pi_k^s \varphi\left((W^{-1}\mathbf{x}_i)_s, \mu_k^s, \nu_k^s\right)\right). \tag{5}$$

The whole IFA parameters $\boldsymbol{\psi}$ can therefore be estimated by maximizing the likelihood function.

2.2.2. Generalized Expectation-Maximization (GEM) Algorithm

When the latent variable densities are known, the unmixing matrix estimation is based on gradient methods that maximize the likelihood. The gradient of the log-likelihood defined in (5) can be derived as:

$$\frac{\partial \mathcal{L}(W; \mathbf{X})}{\partial W} \propto (W^{-1})^t - \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}(W\mathbf{x}_i)\mathbf{x}_i^t, \tag{6}$$

where

$$\mathbf{g}(\mathbf{z}) = \left[\frac{-\partial \log(f^{\mathcal{Z}_1}(z_1))}{\partial z_1}, \ldots, \frac{-\partial \log(f^{\mathcal{Z}_S}(z_S))}{\partial z_S}\right]^t. \tag{7}$$

The update rule of the unmixing matrix is thus given by:

$$W^{(q+1)} = W^{(q)} + \tau\left(\left((W^{(q)})^{-1}\right)^t - \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}(W^{(q)}\,\mathbf{x}_i)\mathbf{x}_i^t\right), \tag{8}$$

where τ is the gradient step that can be adjusted by means of linear search methods ([22]).

The convergence of this algorithm can be improved by using the natural gradient ([16, p. 67, p. 208]).

Maximum likelihood of the whole model parameters can be achieved by an alternating optimization strategy. The gradient algorithm is indeed well suited to optimize the log-likelihood function with respect to the unmixing matrix W when the parameters of the source marginal densities are frozen. Conversely, with W kept fixed, an EM algorithm can be used to optimize the likelihood function with respect to the parameters of each source. These remarks have led use a Generalized EM algorithm (GEM) [23, 24] that simultaneously maximizes the likelihood function with respect to all model parameters.

3. Semi-supervised learning in Independent Factor Analysis

3.1. Derivation of a Generalized Likelihood Criterion

The IFA model is often considered within an unsupervised learning framework. This section considers the learning of this model in a partially-supervised learning context where partial knowledge of the cluster membership of some samples is available. For that purpose, the model is built from a combination of M labeled and $N - M$ unlabeled samples. Consequently, the criterion can be decomposed into two parts corresponding, respectively, to the supervised and unsupervised learning examples and the log-likelihood criterion (5) can be written as:

$$\begin{aligned}\mathcal{L}(W;\mathbf{X}) = N\log(|\det(W|)+ \\ \sum_{i=1}^{M}\sum_{s=1}^{S}\sum_{k=1}^{K_s} l_{ik}^s \log\left(\pi_k^s\varphi\left((W\mathbf{x}_i)_s,\mu_k^s,\nu_k^s\right)\right)+ \\ \sum_{i=M+1}^{N}\sum_{s=1}^{S}\log\left(\sum_{k=1}^{K_s}\pi_k^s\varphi\left((W\mathbf{x}_i)_s,\mu_k^s,\nu_k^s\right)\right). \quad (9)\end{aligned}$$

Note that $l_{ik}^s \in \{0,1\}^{K_s}, l_{ik}^s = 1$ if sample i comes from component c_k of sources s and $l_{ik}^s = 0$ otherwise.

3.2. Practical Considerations

A Generalized EM algorithm (GEM), (also noted here as Algorithm 1) can be designed to simultaneously maximize the likelihood function with respect to all the model parameters [5, 6]. This algorithm is similar to the EM algorithm used to estimate IFA parameter in an unsupervised setting [18], except for the E step, where the posterior probabilities t_{ik}^s are only computed for the unlabeled samples. The score function g of each latent variable density are given by:

$$\begin{aligned} g_s(z_{is}) &= \begin{cases} \frac{-\partial\log\left(\sum_{k=1}^{K_s} l_{ik}^s\pi_k^s\varphi(z_{is};\mu_k^s,\nu_k^s)\right)}{\partial z_{is}}, \text{ if } i \le M \\ \frac{-\partial\log\left(\sum_{k=1}^{K_s} \pi_k^s\varphi(z_{is};\mu_k^s,\nu_k^s)\right)}{\partial z_{is}}, \text{ if } i > M \end{cases} \\ &= \begin{cases} \sum_{k=1}^{K_s} l_{ik}^s \frac{(z_{is}-\mu_k^s)}{\nu_k^s}, \text{ if } i \le M \\ \sum_{k=1}^{K_s} t_{ik}^s \frac{(z_{is}-\mu_k^s)}{\nu_k^s}, \text{ if } i > M \end{cases} \quad (10)\end{aligned}$$

Note that t_{ik}^s is the posterior probability that the sample i belongs to component k of the latent variable s, given $z_{is} = (W\mathbf{x}_i)_s$ and the labels:

$$t_{ik}^s = \frac{\pi_k^s\varphi(z_{is};\mu_k^s,\nu_k^s)}{\sum_{k'=1}^{K_s}\pi_{k'}^s\varphi(z_{is};\mu_{k'}^s,\nu_{k'}^s)}. \quad (11)$$

4. Simulations and discussion

In this section, we investigated the interest of our approach with three real datasets of which the characteristics are given on Table 1. The first dataset is the *Crabs* dataset [1] that concerns the recognition of crabs species and sexes in a population of crabs using different morphological measurements. The second dataset is the well known Fisher's *Iris* data available on-line [2]. The third dataset is the YaleB face database [39]. We choose the first 5 subjects from the dataset and get totally 320 face samples that were captured under different illumination conditions (5 subjects $\times$ 64 illumination conditions).

To better understand our approach as compared to the unsupervised IFA model, different experiments were carried out to show the influence of learning the IFA model when information regarding the component membership of some training samples is introduced. We show that such information can be exploited to efficiently extract a reduced set of variables from the original ones. We show also the potential benefit of incorporating labels in terms of simplification of the optimization problem. This consideration has high practical interest as the problem of local maxima is very important for independent factor analysis.

Three different learning strategies were compared, namely, PCA, unsupervised IFA, and semi-supervised IFA using some labelled samples over all latent variables. The IFA model provides two levels of interpretation corresponding to discrete and continuous latent variables. While results for continuous variables are useful to visualize the projection of the data onto the two-dimensional principal subspace, those of discrete variables allow quantification of classification rates according to each latent variable. In this case, a Maximum a posteriori (MAP) criterion is used to assign each observation to one of the mixture components modeling each latent variable density function. When PCA is applied to reduce dimensionality, the nearest neighborhood (1-NN) classifier is employed for classification.

A few data samples from the whole data are randomly chosen as labeled samples. Twenty random starting points were used for the GEM algorithm and only the best solution according to the likelihood was kept. The performances were quantified using the correct detection rates according to each latent variable calculated on a test set constituted by the samples not labelled during the training phase. The process is repeated for 30 runs and the averaged results are recorded.

[1] http://rweb.stat.umn.edu/R/library/MASS/html/crabs.html
[2] http://mlearn.ics.uci.edu/MLRepository.html

Algorithm 1: Pseudo-code for Semi-supervised IFA with GEM algorithm

Input: Centered observation matrix $\mathbf{X}$, cluster membership for the M labeled data l_{ik}^s

 # Random initialization of IFA parameter vector

$\boldsymbol{\psi}^{(0)}, \quad q = 0$

while *convergence test* **do**

 # latent variable update

$\mathbf{Z} = \mathbf{X}.W^{(q)^t}$

 # Update of the latent variable parameters / EM

forall $s \in \{1, \ldots, S\}$ *and* $k \in \{1, \ldots, K_s\}$ **do**

 # E-Step

$$t_{ik}^{s(q)} = \frac{\pi_k^{s(q)} \varphi(z_{is}; \mu_k^{s(q)}, \nu_k^{s(q)})}{\sum_{k'=1}^{K_s} \pi_{k'}^{s(q)} \varphi(z_{is}; \mu_{k'}^{s(q)}, \nu_{k'}^{s(q)})}, \quad \forall i \in \{1, \ldots, N-M\}$$

$$t_{ik}^{s(q)} = l_{ik}^s, \quad \forall i \in \{1, \ldots, M\}$$

forall $s \in \{1, \ldots, S\}$ *and* $k \in \{1, \ldots, K_s\}$ **do**

 # M-step , Update of the parameter vector of each latent variable

$$\pi_k^{s(q+1)} = \frac{1}{N} \sum_{i=1}^N t_{ik}^{s(q)}$$

$$\mu_k^{s(q+1)} = \frac{1}{\sum_{i=1}^N t_{ik}^{s(q)}} \sum_{i=1}^N t_{ik}^{s(q)} z_{is}$$

$$\nu_k^{s(q+1)} = \frac{1}{\sum_{i=1}^N t_{ik}^{s(q)}} \sum_{i=1}^N t_{ik}^{s(q)} (z_{is} - \mu_k^{s(q+1)})^2$$

 # Update of the score matrix G (10)

$\mathbf{G} = \mathbf{g}^{(q+1)}(\mathbf{Z})$

 # Gradient (8)

$\Delta W = ((\left((W^{(q)})^{-1}\right)^t - \frac{1}{N} \sum_{i=1}^N \mathbf{g}(W^{(q)} \mathbf{x}_i) \mathbf{x}_i^t$

 # Linear search τ (gradient step)

$\tau^* = \text{Linearsearch}(W^{(q)}, \Delta W)$

 # Unmixing matrix update

$W^{(q+1)} = W^{(q)} + \tau^* . \Delta W$

 # Latent variable normalization to remove scale indetermination

forall $s \in \{1, \ldots, S\}$ **do**

$$\sigma_s^2 = \sum_{k=1}^{Ks} \pi_k^{s(q+1)} (\nu_k^{s(q+1)} + \mu_k^{s(q+1)^2}) - \left(\sum_{k=1}^{Ks} \pi_k^{s(q+1)} \mu_k^{s(q+1)}\right)^2$$

forall $k \in \{1, \ldots, K_s\}$ **do**

$$\mu_k^{s(q+1)} = \mu_k^{s(q+1)} / \sigma_s$$

$$\nu_k^{s(q+1)} = \nu_k^{s(q+1)} / \sigma_s^2$$

$$W_{s.}^{(q+1)} = W_{s.}^{(q+1)} / \sigma_s$$

$q \leftarrow q + 1$

Output: Estimated parameters : $\widehat{\boldsymbol{\psi}}$, estimated latent variables : $\widehat{\mathbf{Z}}$

Table 1: Characteristics of real datasets.

name	# dimensions	# samples	# classes
Crabs	5	200	4
Iris	4	150	3
YaleB	40	320	5

4.1. Crabs dataset

The *crabs* dataset consists of 5 morphological measurements recorded for 200 crabs that can be categorized into four groups on the basis of their sex and species: "Blue male", "Blue female", "Orange Male", and "Orange female". The IFA model used to deal with this dataset has 5 latent variables, among which two were modelled by a mixture of 2 normally distribution components and were labelled by using the sex and the species information, the remaining latent variables are modeled by simple Gaussians.

Figure 2 shows the scatter plot of the crabs data when they are projected onto the estimated two-dimensional principal subspace obtained by the semi-supervised IFA model using 20% of labelled training data. As a comparison, the projection of the data onto the first two principal components (PCA) is also given. The semi-supervised IFA leads to a projection giving much better class separation. In fact, the first latent variable clearly captures the variability of the crabs sex while the second corresponds to their species.

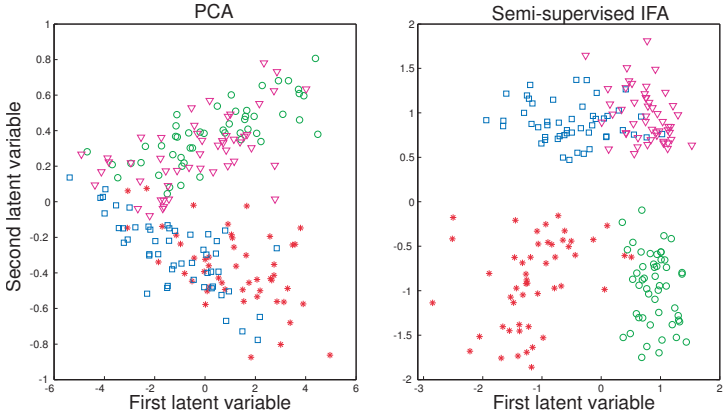

Figure 2: Two-dimensional visualization of the *Crabs* dataset projected onto the first two principal components obtained by PCA and semi-supervised IFA with 20% of labelled samples. In the graph, the stars denote the "blue male" group of crabs, the circles denote the "blue female" group, the boxes indicate the "orange male" crabs and the triangles the "orange female".

In order to accurately quantify the affect of the proportion of labelled samples, the correct detection rates (for the 2 latent variables) has been evaluated as a function of the number of labelled samples. Figure 3 shows the classification performance when the proportion of labelled samples increases from 0 (unsupervised learning case) to 80%. With only 20% of labelled samples, the correct detection rates reach 91.4% for the sex latent variable and 100% for the species latent variable. Note that unsupervised IFA classify correctly classifies 42% of observations for the sex and 56% for the species. PCA-1NN gives

94.8% and 51.8% of correct classification rates for the sex and species. The results are summarized on Table 2.

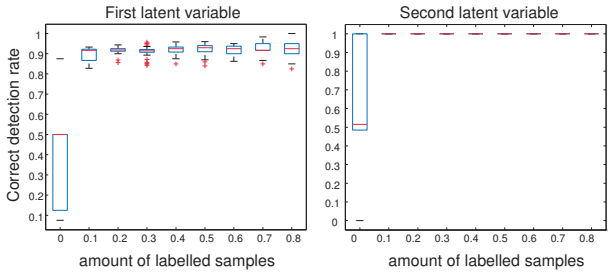

Figure 3: Influence of the proportion of labelled samples on the estimation of the semi-supervised IFA model: Boxplot of the correct detection rates for the first (crab sex) and the second latent (crab species) variables function of the percentage of labelled samples.

Figure 4 displays the CPU time required for the GEM algorithm convergence function of the amount of labelled samples. It can be seen that the time computation (or the number of iterations) exponentially decreases when the amount of labelled samples increases. This graph highlights the potential benefit of incorporating labels in terms of simplification of the optimization problem.

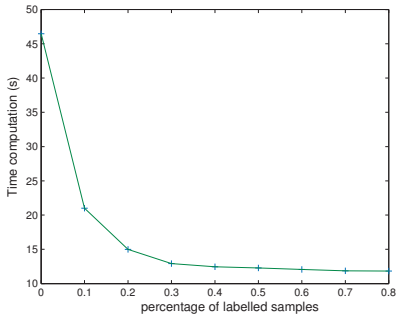

Figure 4: Time computation over 20 random initializations for the GEM algorithm as a function of labelled samples.

4.2. *Iris dataset*

The *Iris* dataset consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). 4 features were measured from each sample : the length and the width of sepal and petal. The IFA model used for this dataset has only one latent variable with a mixture density of three components, (one for each species), the remaining variables being as usual simple Gaussians.

Figure 5 shows the 2D data visualization obtained by the PCA and the semi-supervised IFA with 10% of labelled training samples. It can be seen that semi-supervised IFA requires one latent variable with three components to highlight the latent structure of the *Iris* dataset. The less separable classes are the virginica and the versicolor species while the setosa is the best predictible class. In terms of classification, 86% of samples are well classified in the principal subspace given by the PCA-1NN while only 33% of correct detection rate achieved by unsupervised IFA. This is mainly due to latent variable permutation. By permuting the mixture components a correct detection rate of 97.1% is reached by unsupervised IFA. This finding suggests that the species structure is very clear in this dataset.

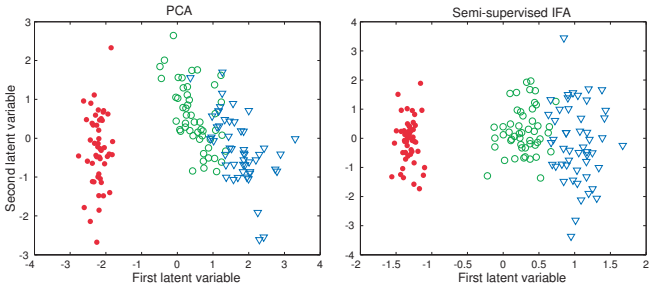

Figure 5: Two-dimensional visualization of the *Iris* dataset projected onto the first two principal components obtained by PCA and semi-supervised IFA with 10% of labelled samples. In the graph, the stars denote the "setosa" group, the circles denote the "versicolor" group and the triangles the *virginica*" species.

Figure 6 shows both the correct detection rate obtained according to the first latent variable and the CPU time computation as functions of the proportion of labelled samples. It can be seen that the performances are drastically improved when 10% of samples are labelled and slows down with more labelled data. This means that too many labelled data are not useful to improve the results for the Iris dataset because as already noted the species structure is very clear in this dataset.

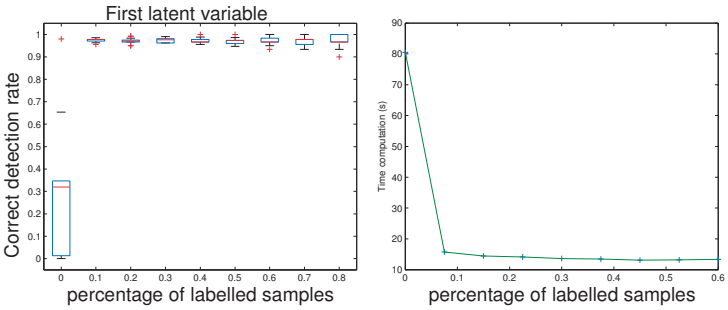

Figure 6: Influence of the proportion of labelled samples on the estimation of the semi-supervised IFA: (a) Boxplot of the correct detection rates evaluated with the MAP criterion on the first latent variable. (b) Time computation over 20 random initializations for the GEM algorithm as a function of the proportion of labelled data

The results are also summarized in Table 2.

4.3. *YaleB face dataset*

5 subjects have been chosen from the YaleB dataset. For each one of them, 64 face samples have been captured under different illumination conditions. The complexity of dimensionality reduction task is due to the lightning conditions that lead to variability between images of the same subject greater than that of different subjects. For this dataset a pre-processing step was necessary to use semi-supervised IFA. In fact considering the pixels gray level as the input variable of the model will lead to a model with 32256 variables one for each pixel. We therefore

first process the data by PCA and kept only the 40 leading principal components. The IFA models was then fixed as follows : 5 latent variables with mixture densities and 35 Gaussian latent variables. Each mixture has 2 components; one component encoding the membership of a picture to a specific subject and the remaining components encoding the membership of the picture to the remaining subjects. Therefore each latent variable with a mixture density can be used to classify picture according to one subject against all the others. Such modeling is partly in contradiction with the hypothesis of Independent Factor Analysis but we will see that good results are still obtained.

When the data are projected onto the first principal subspace given by the PCA, an important overlapping between the 5 groups is noticed. This is illustrated in Figure 7. However, it can be seen on the same figure that semi-supervised IFA with 40% of labelled samples captures much better the intrinsic structure of the dataset. By projecting the dataset on the latent variables corresponding to subjects 1 and 2 one can see that these variables are very good to distinguish these subjects from the others. The complete results averaged on the 5 subjects are summarized on Table 2. Figure 8 shows the boxplot of correct detection rate obtained for each subject when the proportion of labelled samples increases from 0 (unsupervised learning case) to 80%. With 40% of labelled samples, the classification performance reach 87% while it is equal to 50% in the unsupervised IFA.

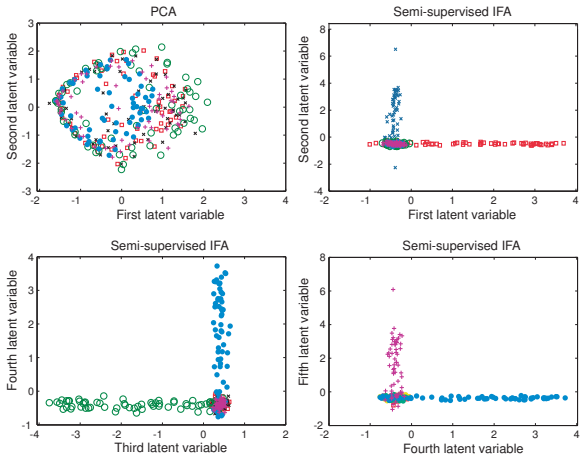

Figure 7: Two-dimensional visualization of the *YaleB* face dataset projected onto the first two principal components obtained by PCA and semi-supervised IFA with 40% of labelled samples. In the graph, different symbols are used to denote the 5 subjects.

5. Conclusion

In this paper, we have proposed a new approach of dimensionality reduction based on partially supervised Independent Factor Analysis. We introduced a criterion where the likelihood of both labelled and unlabelled samples is maximized by a GEM algorithm. Experimental results show that efficient dimensionality reduction can be achieved for some problems where unsupervised methods fail to capture the underlying structure of the data. The amount of labelled data required

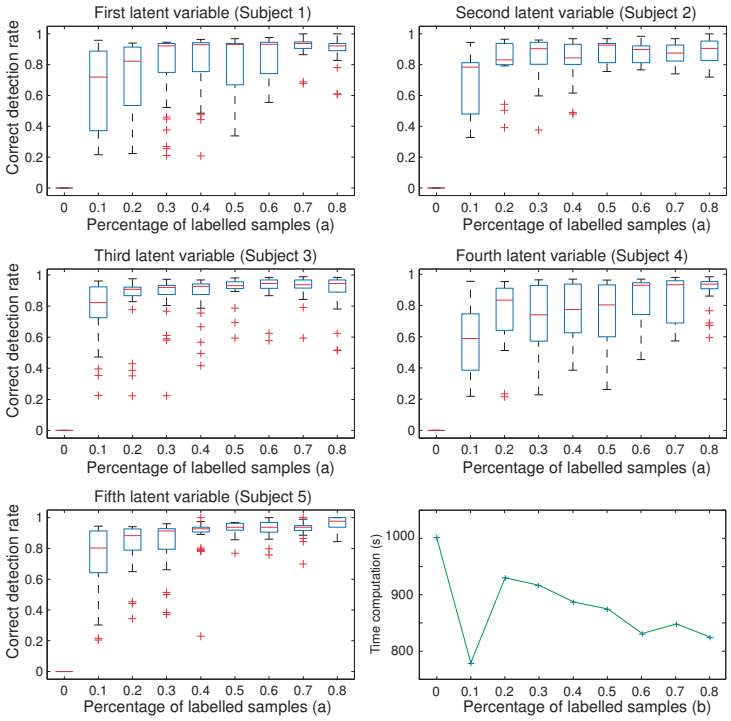

Figure 8: Influence of the proportion of labelled samples on the estimation of the semi-supervised IFA: (a) Boxplot of the correct detection rates evaluated with the MAP criterion on the five latent variables. (b) Mean time computation over 20 random initializations for the GEM algorithm as a function of the proportion of labelled data

to reach a satisfactory accuracy depends on the discrimination problem complexity but for all the problems investigated here we find that with only 10% of labelled data semi-supervised IFA gives results much better than PCA and ordinary IFA.

A. Appendix : Gradient of the Log-likelihood with respect to the unmixing matrix

The log-likelihood in the noiseless IFA is given by:

$$\mathcal{L}(\boldsymbol{\psi};\mathbf{X}) = \sum_{i=1}^{N}\sum_{s=1}^{S} \log\left(f^{\mathcal{Z}_s}\left((W\mathbf{x}_i)_s\right)\right) + N\log(|\det(W)|),$$

In order to compute the gradient of $\mathcal{L}(\boldsymbol{\psi};\mathbf{X})$ with respect to W, we have to compute the derivative of the logarithm of the absolute value of a matrix determinant with respect to one of its elements, which is given by (see [4], [3, p. 8]):

$$\frac{\partial \log(|\det(X)|)}{\partial X_{lk}} = (X^{-1})_{kl},$$

By using this relationship and assuming that latent variable densities $f^{\mathcal{Z}_1},\ldots,f^{\mathcal{Z}_S}$ are known, the derivative of the log-likelihood with respect to one element W_{lk} of the unmixing matrix can be written as:

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\psi};\mathbf{X})}{\partial W_{lk}} &= N(W^{-1})_{kl} + \sum_{i=1}^{N}\frac{\partial \log\left(f^{\mathcal{Z}_l}\left((W\mathbf{x}_i)_l\right)\right)}{\partial W_{lk}} \\ &= N(W^{-1})_{kl} - \sum_{i=1}^{N}\mathbf{x}_{ik}g_l\left((W\mathbf{x}_i)_l\right),\end{aligned}$$

where $g_l(z)$ is the opposite of the derivative of the logarithm of the latent variable density z_l :

$$g_l(z) = \frac{-\partial \log(f^{\mathcal{Z}_l}(z))}{\partial z}$$

By using matricial notations, we can define the $\mathbf{g}$ function:

$$\begin{aligned} \mathbf{g} &: \mathbb{R}^S \rightarrow \mathbb{R}^S \\ \mathbf{g(z)} &= \left[\frac{-\partial \log(f^{\mathcal{Z}_1}(z_1))}{\partial z_1}, \ldots, \frac{-\partial \log(f^{\mathcal{Z}_S}(z_S))}{\partial z_S} \right]^t. \end{aligned}$$

Which allows us to obtain the matrix of the derivative of the log-likelihood with respect to each element of W :

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{\psi}; \mathbf{X})}{\partial W} &= N(W^{-1})^t - \sum_{i=1}^{N} \mathbf{g}\left(W\mathbf{x}_i\right) \mathbf{x}_i^t \\ &\propto (W^{-1})^t - \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}\left(W\mathbf{x}_i\right) \mathbf{x}_i^t. \end{aligned}$$

References

[1] I.T. Jolliffe. Principal Component Analysis. Springer-Verlag, 1986.

[2] J.E. Jackson. A User's Guide to Principal Components. New York: John Wiley and Sons, 1991.

[3] K.B. Petersen and M.S. Pedersen. The Matrix Cookbook. 2008.

[4] D. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. 1996.

[5] E. Côme and L. Oukhellou and T. Denœux and P. Aknin. Partially-supervised learning in Independent Factor Analysis In 17th *Proceedings of the European Symposium of Artificial Neural Network (ESANN 09)*, 2009.

[6] E. Côme and L. Oukhellou and T. Denœux and P. Aknin. Noiseless Independent Factor Analysis with mixing constraints in a semi-supervised framework. Application to railway device fault diagnosis. In 19th *International Conference on Artificial Neural Networks (ICANN 09)*, 14-17 September 2009, Limassol, Cyprus, LNCS 5769, pp 416-425, Springer-Verlag.

[7] E. Côme. Apprentissage de modèles génératifs pour le diagnostic de systèmes complexes avec labellisation douce et contraintes spatiales. PhD thesis, Université de Technologie de Compiègne, 2009.

[8] T. Cox and M. Cox. Multidimensional Scaling. Chapman and Hall,London, 1994.

[9] L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha and D. D. Lee. Spectral methods for dimensionality reduction. In Semi-Supervised Learning (O. Chapelle, B. Schölkopf and A. Zien, eds.). MIT Press, pp. 293-308. 2006

[10] C. J. C. Burges. Geometric methods for feature extraction and dimensional reduction. Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Kluwer Academic Publishers, 2005.

[11] K. Fukunaga. Introduction to statistical pattern recognition, 2nd edition, Academic Press, New York, 1990.

[12] J.H. Friedman. Exploratory projection pursuit. Journal of the American Statistical Association, 82(397):249–266, 1987.

[13] M.C. Jones and R. Sibson. What is projection pursuit ? Journal of the Royal Statistical Society, ser. A, 150:1–36, 1987.

[14] W. B. Johson and J. Lindenstrauss. Extensions of Lipschits Mapping into a Hilbert Space. Contemp. Math., 26, pp. 189–206,1984.

[15] T. Watanabe, E. Takimoto and A. Maruoka. Dimensionality Reduction by Random Projection. (in Japanese), Tech. Rep. of IEICE, COMP 2001-92, 2002.

[16] A. Hyvärinen and Juha Karhunen and Erkki Oja. *Independent Component Analysis*. Wiley, 2001.

[17] A. J. Bell and T. J. Sejnowski An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[18] H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.

[19] E. Moulines, J. Cardoso, E. Cassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3617–3620, 1997.

[20] A. Montanari, D. G. Calo, C. Viroli. Independent factor discriminant analysis. Computational Statistics & Data Analysis 52(6): 3246–3254 (2008)

[21] T. Bakir, A. Peter, R. Riley, and J. Hackett. Non-negative maximum likelihood ICA for blind source separation of images and signals with application to hyperspectral image subpixel demixing. In *Proceedings of the IEEE International Conference on Image Processing*, pages 3237–3240, 2006.

[22] J. Nocedal and S. J. Wright. Numerical Optimization. Springer Series in Operations Research, Springer, 1999.

[23] A.P. Dempster and N.M. Laird and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, B(39):1–38, 1977.

[24] G. J. Mclachlan and T. Krishnan. The EM algorithm and Extension. Wiley, 1996.

[25] G. Shafer. A mathematical theory of evidence. *Princeton University Press*, 1976.

[26] E. Côme, L. Oukhellou, T. Denœux and P. Aknin. Learning from partially supervised data using mixture models and belief functions. *Pattern recognition*, 42:334–348, 2009.

[27] C. Spearman. General intelligence, objectively determined and measured. *American Journal of psychology*,15:201–293, 1904.

[28] L. L. Thusrstone. Multiple Factor Analysis. University of Chicago Press, 1947.

[29] S. Amari and A. Cichocki and H. H. Yang. A New Learning Algorithm for Blind Signal Separation. Proceedings of the 8th Conference on Advances in Neural Information Processing Systems (NIPS). 8:757–763, MIT Press, 1996.

[30] D. J. Bartholomew and K. Martin. Latent variable models and factor analysis. Kendall's library of satistics, Seconde édition, Arnold, London, 1999.

[31] B. S. Everitt. An Introduction to Latent Variable Models. Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York, 1984.

[32] S. Kaski. Data exploration using self-organizing maps. PhD thesis, Helsinki University of Technology,Finland, 1997.

[33] S. Kaski. Dimensionality reduction by random mapping: fast similarity computation for clustering. Proc. IEEE International Joint Conference on Neural Networks, 1:413–418, 1998.

[34] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing*. Wiley, 2002.

[35] T.K. Kohonen. The self-organizing map. Proc. IEEE, 78:1484–1480, 1990.

[36] M.A. Carreira-Perpinan. A review of dimension reduction techniques. Technical report CS-96-09. Department of Computer Science, University of Shefield, 1997.

[37] N. Kwak, CH. Kim, H. Kim. Dimensionality reduction based on ICA for regression problems. Neurocomputing, 71, 13-15:2596–2603, Elsevier, 2008.

[38] N. Kwak, C.H. Choi. Feature extraction based on ICA for binary classification problems. IEEE Trans. Know. Data. Eng 15, (6):1374–1388, 2003.

[39] A.S. Georghiades, and P.N. Belhumeur, and D.J. Kriegman. From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. IEEE Trans. Pattern Anal. Mach. Intelligence 23, (6):643–660, 2001.

Table 2: Correct detection rates (%) averaged over 30 independent datasets, for three dimensionality reduction methods PCA-1NN, IFA and semi-supervised IFA with different proportions of labelled samples (%).

	PCA-1NN	IFA	Semi-supervised IFA						
	0	0	*10*	*20*	*30*	*40*	*50*	*60*	*70*
Crabs	55.8	49.5	94.7	95.7	95.4	95.8	96	95.8	96.2
Iris	86.1	33.1	97.4	97.1	97.5	97.3	97.1	97.1	97.1
YaleB	25.8	49.7	71	75.7	80.1	84.1	86.9	87.5	89